DOCUMENT RESUME

ED 253 574 TM 850 118

AUTHOR Sanders, James R.
TITLE Reflections on Evaluation Plans for Bilingual Projects.
PUB DATE [81]
NOTE 18p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bilingual Education Programs; Compensatory Education; *Data Collection; Elementary Secondary Education; *Evaluation Methods; Evaluation Needs; Federal Aid; Government School Relationship; Program Implementation; Research Design; Testing Problems
IDENTIFIERS Elementary Secondary Education Act Title I; *Elementary Secondary Education Act Title VII; *Evaluation Problems

ABSTRACT

Many difficulties are encountered when large-scale monitoring of federally-funded programs (including Title VII bilingual education programs) has been attempted. These difficulties may be categorized as follows: (1) differences in perception of the purpose of monitoring; (2) variance in the definition of the treatment (program) and expectations for results; (3) measurement issues and difficulties; (4) design issues and difficulties; and (5) cooperation and control in data collection. After discussing each of these areas in terms of design considerations for a system to monitor Title VII projects, this paper concludes with a summary of thoughts and recommendations about a strategy for collecting required information about bilingual education projects. A 33-item bibliography is included. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED253574

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

REFLECTIONS ON EVALUATION PLANS FOR BILINGUAL PROJECTS

© James R. Sanders
Evaluation Center
Western Michigan University

Large-scale federal monitoring of compensatory educational programs has received the attention of a large number of people over the past fifteen years. These people have played different roles in the delivery and monitoring of these programs. Some have been congressional aides (Cross, 1979), some researchers in independent R & D agencies (Gamel, et al., 1975, Wargo, et al., 1972), some researchers in federal government agencies (Hill, et al., 1978), some university-affiliated researchers (Linn, 1979; Jaeger, 1979), some federal, state, and local program staff (Anderson, et al., 1978; Wisler and Anderson, 1979), and some part of a network of technical assistance providers set up by the federal government (Storlie, et al., 1978; Hanson, 1978; Echternacht, 1978). Although most of this attention has been on ESEA, Title I program monitoring, much of the experience gained in the recent past can be transferred to the monitoring of bilingual program accomplishments.

One reason for the amount of attention given to federal monitoring of compensatory educational programs has been the number of difficulties encountered when large-scale monitoring has been attempted. These difficulties may be categorized as follows:

1. Differences in perception of the purpose of monitoring
2. Variance in the definition of the treatment (program) and expectations for results
3. Measurement issues and difficulties
4. Design issues and difficulties
5. Cooperation and control in data collection

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. R. Sanders
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 118

This discussion of design considerations for a system to monitor Title VII bilingual education projects is organized by the five topics listed above. A concluding section of this position paper contains a summary of thoughts and recommendations about a strategy for collecting required information about bilingual education projects.

Differences in Perceptions of the Purpose of Monitoring

A performance review of the ESEA, Title I Technical Assistance Centers (Millman, et al, 1979) made the distinction between reporting systems and evaluation systems in collecting information on federally funded projects. Reporting systems provide selected summative information about the project to address a specific Federal need for information. Evaluation systems provide information specific to an individual project that may be used at some level (local, state, or federal) to make project improvements or acccuntability statements about the particular project. The former is narrow in the scope of information being collected and emphasizes the collection of information which is comparable across all projects and which may be aggregated over all projects to provide a summary statement about the collection of projects. The latter is wide in the scope of information being collected and is tailored to the idiosyncratic nature of a particular project. The results of an evaluation system are revealing about one project, but may not be able to be used for comparisons or aggregation.

Webster and Stufflebeam (1980) have made a similar distinction between quasi-evaluations and true evaluations. Quasi-evaluation studies begin with a particular question and then move to the methodology appropriate for answering that question. Examples of such studies include objectives-based studies, accountability studies, experimental research studies, testing

programs, and management information systems. True evaluations, on the other hand, are designed and conducted to assist some audience(s) to judge and improve the value of a particular program or project. Examples of such studies include accreditation/certification studies, policy studies, decision oriented studies, consumer oriented studies, client-centered studies, and connoisseur-based studies of a program or project.

Another distinction that is pertinent to the discussion of purpose for a system was developed by Stufflebeam, et al (1971). They noted that studies can serve decision making at different levels and, "Only chance renders an information system adequate when the decision maker does not specify the expected decision situations to be served by the system." (Stufflebeam, et al, 1971, p. 123). Information collected for a particular purpose at one level may have little or no utility at other levels of a national system of projects.

The present purpose of gathering information on Title VII bilingual education projects, as I understand it, is to provide evidence of impact across funded projects to Congressional decision makers. In particular, the need is to document gains in English proficiency by those students who have been involved in bilingual projects and to provide evidence that such gains cannot be attributed to normal growth in regular educational programs alone.

Hence, the purpose of gathering information best fits the concepts of a reporting system and a question-oriented study as developed by Millman, et al., (1979) and Webster and Stufflebeam (1980) respectively. It will be wise to avoid the label of "evaluation" for the system being developed so that the criteria used for judging the system will be appropriate for the actual purpose of the system. Furthermore, it will be important to keep in mind the level of decision making for which the system is being developed. Since it is to efficiently and effectively serve Congressional decision makers,

the system should not be expected to provide information pertinent to the needs of local project and state decision makers.

Variance in the Definition of the Treatment (Program) and Expectations for Results

Consistent with the principle of local educational autonomy that prevails in this country, local projects are typically defined to meet the particular needs of the students being served. This phenomenon was found in studies of ESEA, Title I and Follow Through projects (David, 1978; House, et al., 1978) and there is no reason why it shouldn't be the case for Title VII bilingual education projects. If this is the case, then individual project objectives and educational activities will vary considerably across the country. This reality, then, has implications for the design of a system that provides information that can be aggregated across projects and also allows for comparisons to be made across projects.

Two implications for the design of an information gathering system emerge from this situation. First, a wide range of project objectives must be accommodated so that mismatchs with individual project objectives can be avoided. Second, process information about the activities of each individual project will be required so that projects may be grouped appropriately for comparisons and aggregation.

One way to approach the first requirement is to develop a frame of objectives, and associated measurement devices for each (test items, interview questions, performance exercises, attitude or interest scales, social development scales, personal growth scales). This frame may then be used to select measures that match the objectives of individual projects. Over a large number of projects, a picture of student development may then be built that does accommodate a variety of locally defined goals. A similar design was

discussed by Stake (1971a) in his analysis of performance contracting in education.

A strategy to approach the second requirement is to develop a standard project description form which will provide information about the activities of the project. Information about curriculum content, teaching methods, and time spent on each objective are three project characteristics that can be expected to vary. There will be others and well-planned background case studies of Title VII bilingual education projects may be used to prepare for the development of a project information form.

Measurement Issues and Difficulties

There are two measurement issues that have emerged from previous large-scale attempts at gathering comparable student performance information that are important to consider. One difficulty has been with a lack of conceptual clarity about what should be measured. Large-scale studies have tended to focus on stated project objectives in order to select measurement instruments. Experiences with local projects have shown, however, that the actual objectives and consequences of a project tend to drift away from stated objectives as the project evolves (e.g., House, et al., 1978). What is written and what is done differ and the design for monitoring Title VII bilingual education projects should consider the latter. As previously noted, background case studies could help in the development and selection of appropriate measurement instruments.

The second issue has been with the use of standardized tests for gathering data about the impact of compensatory educational projects on students. Such test scores have not been sensitive to change (Stake, 1971a). They frequently have not provided a good match with actual project outcomes (Stake, 1971 a,b; House, et al., 1978), and they are highly corruptible (Stake, 1971a).

Regarding the insensitivity to change, Stake (1971a) noted:

> Most tests are indirect measures of educational gains, correlates of learning rather than direct evidence of achievement. Correlation with important general learning is often high, but correlation of test scores with performance on many specific educational tasks is seldom high...the test developer's basis for improving tests has been to work toward better predictions of later performance rather than better measurement of present performance. Assessment of what a student is now capable of doing is not the purpose of most standardized tests. (Stake, 1971a, p. 3)

The difficulty of the mismatch between what standardized test measure and the actual outcomes of funded projects can be traced to the professional practice of project staff in providing instruction and educational experiences that are needed at the moment by their students, rather than mechancially teaching a standard curriculum. While standardized tests are intended to provide survey coverage of the central tendency of content in a subject, they do not provide adequate concentration on any specific goal areas and they do not represent the purposes of instruction to a special subgroup of the norm population (these students who are in bilingual projects). Finally, there is substantial evidence that use of standardized tests as a criterion measure in project monitoring affects the focus of instruction. Teachers will teach for the test, making it the curriculum, regardless of the needs of individual students. Thus, standarized test data as evidence of student learning becomes instead evidence of the teacher's willingness to suspend teaching that will be tailored to his or her students in favor of providing instruction and experiences that will affect students' test scores. All of this suggests that measurements need to be tailored to individal projects while at the same time they need to provide comparability across projects. The use of a broad frame of measurement devices and items, from which project examiners may select those that are appropriate to actual instruction directed toward individual student needs

is a measurement goal that should be incorporated into the monitoring of Title VII bilingual education projects.

Design Issues and Difficulties

There is a tremendous amount of literature on the design of wide-scale studies to assess the impact of federal programs (Campbell, 1976; Boruch and Wortman, 1979; Riecken and Boruch , 1974; Bennett and Lunsdaine, 1975;), on the design of particular large-scale reporting systems (e.g., Anderson and Wisler, 1979; Tallmadge and Wood, 1978), and on the assessment of group change (Harris, 1963; Linn and Slinde, 1977). Millman, et al., (1979) compiled a list of issues related to the ESEA, Title I reporting models, while Linn (1978, 1979) and Jaeger (1979) provided still further discussion of the Title I approach to gathering national impact data. This literature suggests several criteria and principles which could provide a structure for the monitoring of Title VII bilingual education projects.

One criterion is that the data gathered on students participating in bilingual education projects must be comparable and able to be aggregated. This requirement is necessary if the reporting system is to meet the expectations of those who have mandated this system. Jaeger (1979) suggested that in its simplest form, descriptors like "highly effective," "effective", "ineffective", and "aversive" for individual projects, based on a standard set of criteria for each category, would meet the letter of the mandate. While this approach to providing comparable and aggregated evidence of impact would avoid the technical weakness of the Title I models (Linn, 1979; Jaeger, 1979; Millman, et al., 1979), it may not provide sufficient information about student development for the audiences of the reporting system.

A second criterion is that multiple measurements on participating students is needed in order to report on change. In its simplest form, pretest and posttest information in one school year would meet this criterion. However, two measurements will be insufficient to detect changes that are gradually introduced over the year (or over several years for some students). Time series (Glass, Willson, and Gottman, 1972) longitudinal, and time lag designs (Goulet, 1975) are workable single group design alternatives to the design approaches used in Title I.

Linn and Slinde (1977) reviewed several approaches to assessing pre- and posttest change and concluded (as did Cronbach and Furby, 1970) that, "Problems in measuring change abound and the virtues in doing so are hard to find" (Linn and Slinde, 1977, p. 147). Simple t-tests between group pre-test and post-test scores are defensible from a design standpoint, but rarely will this design detect the gradual changes that will occur in a bilingual education project. Single-group designs with multiple data points may be a workable alternative to monitoring these projects.

A third criterion is that the system design must be workable in terms of logistics, time requirements (i.e., taking students and project staff away from the tasks of education, demands for examiners), cost, and lay understanding of the design. Millman, et al., (1979) found that one of the most frequent concerns of state and local educators about Title I project monitoring was the increased demand for data gathering and reporting imposed by the system. Since the decision making level for the bilingual education project reporting system is the federal level, sampling approaches to data collection may be used to limit costs and time requirements. In addition, the technical assistance network set up by the Title I

reporting system may provide a mechanism for handling the logistics and lay indoctrination into the system. This possibility is one that should be explored.

A fourth criterion is that the system should provide a basis for drawing justifiable conclusions about the impact of Title VII bilingual education projects. This is the sine qua non of the design task. There exist debates over the appropriateness, and credibility, of using quantitative versus qualitative ways of studying impact (Campbell, 1976; Sanders, 1980), which spill over to requirements that data must be empirical or statistical and replicable versus requirements that data must be descriptive of, and responsive to, the reality of project outcomes and changes in the project (treatment) during the course of providing compensatory services. Proponents of quantitative quasi-experimental designs argue that the results of qualitative studies are intrinsically equivocal for making causal inferences mainly because of the confounding of selection and treatment (Campbell, 1976). Proponents of qualitative and naturalistic designs argue that the results of input-output quasi-experimental studies are uninterpretable due to the idiosyncratic nature of projects (treatments) in uncontrolled settings (Campbell, 1976; Guba, 1978). Both sets of arguments assume that the design decision is an either/or decision and that a commitment to one way of knowing precludes using the other. In fact, the decision is not constrained by any methodological principles which preclude designs that make full use of both ways of knowing. The two meta-approaches to designing a reporting system may be used to complement each other. One way to address the development of a design is to consider it as a multi-stage process. The first question to consider is, "What do we know about Title VII bilingual education projects?"

Is there some understanding about the forms that these projects take? Are there patterns in their evolutionary development? in their content? in their methods? in their time allocated to certain training or educational objectives? in the students being served? Such knowledge will serve as a guide to identifying what should be measured and how to measure it. Although a small number of first hand case studies of projects will be revealing and serve as a heuristic for generating a set of working hypotheses, they will not be sufficient for structuring large-scale data collection. Findings from case studies may be followed by a survey of Title VII projects to discover the extent to which certain project characteristics exist and to identify further distinctions that must be incorporated into the reporting system. This first stage of the design could be considered the conceptual analysis stage. The stage that follows could be considered a pilot stage. At this point in the design process, the best estimate of a design for generating data that will lead to justifiable conclusions about the impact of projects could be implemented on a limited basis. Selecting one state, for example, to pilot test the design and to validate its findings would allow an evaluation of the design to take place. The investment that will ultimately be made in the national reporting system will be substantial and the decisions that will be made using the reports generated by the system will not be inconsequential for bilingual education in this country. Such a situation warrants early evaluation of the design for the reporting system. The final stage, then, could be full scale implementation of the design and continuous monitoring of its operation for the purposes of quality control.

My understanding of the requirements placed on the final design by federal staff is that it must provide for the collection of quantitative evidence of project impact. This evidence must be aggregated across projects and allow for comparisons of results (e.g., over time, across projects). One way to address the criticisms of input - output quasi-experimental designs found in Millman et al., (1979), Linn and Slinde (1977), Jaeger (1979), and Linn (1979) would be to conceive of a single group design with multiple measurements of the same students over time and a sampling plan that will allow for several different types of comparison (longitudinal, cross-sectional, time lag). Such a design has been described by Goulet (1975). While a design approach of this type will provide a basis for separating change due to general education treatment and that due to maturation (chronological age), small-scale studies over a relatively few projects will enable selected hypotheses to be tested in order to answer other questions about the causes of change.

Cooperation and Control in Data Collection

A serious difficulty with the ESEA, Title I reporting system as discussed by Millman, et al, (1979), was in gaining cooperation of local and state staff in implementing the system and in quality control of the data being generated. The cooperation issue was based in part on the amount of time local and state staff would have to spend on testing, data analysis, and report preparation. The quality control issue was in large part due to the dependence on local and state personnel for data collection, data processing, and reporting. There was evidence that errors were introduced from several difference sources.

The ESEA, Title I system initially required all projects to participate and report, although a sampling plan was subsequently introduced. This requirement far exceeds what would be necessary if the purpose of the system is to provide reporting information for federal level decision making. The Title VII bilingual education reporting system may not require such a massive effort and some degree of cooperation and control may be gained through project sampling. Sample sizes for categories of projects defined through back ground (stage 1) studies would have to be determined at some later point in time.

The ESEA, Title I Technical Assistance Centers (TACs) are a resource that may prove to be useful for the bilingual education reporting system. If project sampling is to be used, supplemental contracts to the TACs may provide for a cadre of examiners who would: (1) select from a measurement frame the items and other measurement devices to be used at a particular project site; (2) select the students on whom data will be gathered; (3) gather data; (4) analyze data; and (5) report data according to design specifications. By organizing large-scale data collection in this way, the primary contractor and federal staff will not only control the quality of the data, but will also be able to address the criteria of logistics and need for lay understanding of the design. Moreover, the TACs have already developed a rapport and identity with local and state educators (Millman, et al., 1979) that could facilitate cooperation. Cooperation would also be facilitated if external examiners are used by the absence of demand on local and state staff for time to be spent on data gathering, analysis, and reporting functions.

Concluding Summary: Thoughts About a Strategy for Collecting Required Information

The preceding reflections about a reporting system for Title VII bilingual education projects lead to a set of suggestions for designing the system. These suggestions may be summarized as follows:

1. The design requirements are consistent with previous conceptualizations of reporting systems and do not fit contemporary conceptualizations of program evaluation. The selection of a design should be made with criteria for a reporting system in mind, rather than according to criteria for program or project evaluation (Joint Committee on Evaluation Standards, 1981). Conceptual difficulties can be avoided by avoiding the use of "evaluation" as a label for the system.
2. Differences across local project objectives and accomplishments may require using a frame of objectives and associated measurement devices in order to accommodate a range of potential impacts. Related suggestions have been made by Stake (1971a), and earlier by Cronbach (1963).
3. Process and demographic information about local projects should be gathered as part of the design. This information may be used for categorizing projects (for comparisons and aggregation) and for analysis of project impact (consistent with design plans suggested by Dyer (1970) or Marco (1974), perhaps).
4. Outcome measures on students should be selected based on reality rather than written project objectives. Designs which have not considered the actual accomplishments of projects have been severely criticized (House, et al., 1978).

5. Standardized tests should be cautiously scrutinized for their utility and weaknesses in documenting gains in bilingual education projects. The use of a measurement frame from which items and other measurement devices may be selected is a viable design alternative.
6. The problems involved in using gain scores and measuring change (Linn and Slinde, 1977; Cronbach and Furby, 1970) suggest that alternative design conceptualizations be considered. Single-group designs, such as those described by Goulet (1975) will provide important information about students and will allow for small-scale research on hypotheses, testing plausible causes for student outcomes.
7. A multi-stage design development process wherein both qualitative and quantitative studies provide a logical and systematic basis for design decisons may be warranted, depending on how much is known about bilingual education projects in this country.
8. Cooperation and control in implementing the design may be accomplished through the use of a sampling plan and field examiners. The ESEA, Title I Technical Assistance Centers provide an available resource for implementing a national reporting system.

REFERENCES

Anderson, J. K., Johnson, R. T., Fishbein, R., Stonehill, R. M., and Burnes, J. C. The U.S. Office of Education models to evaluate ESEA, Title I: Experiences after one year. Washington, D.C.: Office of Education, 1978.

Bennett, C. A. and Lumsdaine, A. A. (Eds.). Evaluation and Experiment. New York: Academic Press, 1975.

Boruch, R. F. and Wortman, P. M. Implications of Educational Evaluation for Evaluation Policy. In D. C. Berliner (Ed.) Review of Research in Education, Vol. 7. Washington, D.C.: American Educational Research Association, 1979.

Campbell, D. T. Assessing the Impact of Planned Social Change. Kalamazoo, Michigan: Evaluation Center, Western Michigan University, Occasional Paper No. 8, 1976.

Cronbach, L. J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Cronbach, L. J. and Furby, L. How we should measure "change" - or should we? Psychological Bulletin, 1970, 74, 68-80.

Cross, C. T. Title I Evaluation - A case study in Congressional frustration. Educational Evaluation and Policy Analysis, 1979, 1, 15-21.

David, J. L. Local Uses of Title I Evaluations. Final report on HEW Contract HEW-100-77-0095. Menlo Park, California: SRI International, July, 1978.

Dyer, H. S. Toward objective criteria of professional accountability in the schools of New York City. Phi Delta Kappan, 1970, 52, 206-211.

Echternacht, G. The use of different models in the ESEA Title I evaluation system. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, 1978.

Gamel, N. N., Tallmadge, G. K., Wood, C. T., and Binkley, J. L. State ESEA Title I reports: Review and analyses of post reports, and development of a model reporting system and format. Mountain View, California: RMC Research Corporation, 1975.

Glass, G. V, Willson, V.L., and Gottman, J. M. Design and Analysis of Time--Series Experiments. Boulder, Colorado: Laboratory of Educational Research, University of Colorado, 1972.

Guba, E. G. Toward a Methodology of Naturalistic Inquiry in Educational Evaluation. Los Angeles, California: Center for the Study of Evaluation, UCLA, 1979.

Hanson, J. B. Measurement problems in Title I evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, 1978.

Harris, C. W. (Ed.) Problems in Measuring Change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Hill, P., et al., Compensatory Education Study. Washington, D.C.: National Institute of Education, 1978.

House, E. R., Glass, Gene V, McLean, L. D., and Walker, Decker, F. No simple answer: Critique of the Follow Through Evaluation. In T. D. Cook, et al., (Eds.). Evaluation Studies Review Annual, Vol. 3. Beverly Hills, California: Sage Publications, 1978.

Jaeger, R. M. The effect of test selection on Title I project impact. Educational Evaluation and Policy Analysis, 1979, 1, 33-40.

Joint Committee on Standards for Educational Evaluation. Standards for Evaluations of Educational Programs, Projects, and Materials. New York: McGraw-Hill, 1981.

Linn, R. L. Validity of inferences based on the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, 1979, 1, 23-32.

Linn, R. L. and Slinde, J. A. The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 1977, 47, 121-150.

Marco, G. L. A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 1974, 11, 225-234.

Millman, J., Paisley, W., Rogers, W. T., Sanders, J. R., and Womer, F. B. Performance Review of USOE's ESEA Title I Evaluation Technical Assistance Program. Final Report on HEW Contract HEW-300-78-0355. Washington, D.C.: Hope Associates, Inc., April 1979.

Riecken, H. W. and Boruch, R. F. Social Experimentation. A method for Planning and Evaluating Social Intervention. New York: Academic Press, 1974.

Sanders, J. R. Using multiple approaches in field research. Evaluation News, No. 16, Summer, 1980.

Stake, R. E. Measuring what learners learn. Urbana, Illinois: Center for Instructional Research and Curriculum Evaluation, University of Illinois, 1971a.

Stake, R. E. Testing hazards in performance contracting. Phi Delta Kappan, 1971b, 52, 583-589.

Storlie, T., et al. Revised RMC materials committee recommendations. Evanston, Illinois: Educational Testing Service, August 1978.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., and Provus, M. M. Educational Evaluation and Decision-Making. Itasca, Illinois: Peacock, 1971.

Stufflebeam, D. L. and Webster, W. J. An analysis of alternative approaches to evaluation. Educational Evaluation and Policy Analysis, 1980, 2, 5-20.

Tallmadge, G. K. and Wood, C. T. Users Guide. Mountain View, California: RMC Research Corporation, 1978.

Wargo, M. J., Tallmadge, G. K., Michaels, D. D., Lipe, D., and Morris, S. J. ESEA Title I: A reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970. Palo Alto, California: American Institutes for Research, March, 1972.

Wisler, C. E. and Anderson, J. K. Designing a Title I evaluation system to meet legislative requirements. Educational Evaluation and Policy Analysis, 1979, 1, 47-55.